B. Medical Relevance and Collaboration

Our research is being done in collaboration with Professors Raymond Kahn, Theodore Fischer, and William Burkel of the Department of Anatomy, the University of Michigan Medical School. During the three years that we have been in contact, they have been working to define the factors that are necessary for the long-term (9 days or longer) maintenance in vitro of prostate and lung explants. The results of their efforts will enable other researchers to maintain organs while studying the etiology of various diseases affecting these organs or to test the direct action of drugs and hormones. Separate procedures must be developed for each organ type since the maintenance requirements are different for each organ. The cultivation of prostate explants was funded by an NIH contract. Such a technique would be useful for studying the etiology of benign hyperplasia and carcinoma of the human prostate. The cultivation of lung is funded by an NIH grant and is done in collaboration with Dr. Paul Weinholdt, a biochemist at the Veterans Administration Hospital in Ann Arbor. He is studying the formation of surfactant by lung tissue, the absence of which is a primary cause of death in premature infants.

These researchers have found computer methods for the storage and retrieval of their experimental results to be very helpful They are currently using standard database and statistical packages on the University of Michigan's computer system. During one year they perform experiments using approximately 2,000 explants and record values of numerous dependent and independent variables for each one. These data must be analyzed promptly so that new experiments can be planned, based on the results of previous ones. The further power that artificial intelligence techniques could provide would enable them to perform their work more effectively. Our system should relieve investigators from the difficult tasks of trying to reduce multiple variables into a single coded value and trying to give these values consistent meaning across investigators, a recurring problem in this type of analysis. The AI system that we envision will be able to store more information, the kind that is not easily codified for use in their current statistical routines, such as graphic information. Our aim is to build a system that would be able to process their data in a more sophisticated manner, being able to perform analyses and make judgments that hitherto had to be done by the researchers themselves. This should allow the researchers to spend more time actually performing experiments and thinking about basic, theoretical questions. Like other current computer programs, the AI system will be able to perform functions that were previously impossible or impractical due to the vast amount of data that need to be analyzed or to the fine discriminations that need to be made which are beyond human perception. By using natural language input and output, the AI system should be faster and more comfortable to use.

The histologists' role in this project has been to acquaint us with their existing procedures, the types of knowledge they use, and the decisions they make in order to perform organ culture research. We have been attending their weekly progress meetings and interviewing them individually. In addition to developing our AI system, we have given them consultation with regard to the use of existing computer programs on the University of Michigan computer system and the purchase of computer hardware for their department.

C. Progress Summary

Our main emphasis so far has been on the natural language processing aspects of this project. We have written programs using Interlisp at SUMEX that read typed, natural language input and perform a morphological analysis of the words. Due to the large number of Latin words that are used in anatomy, we have defined both English and Latin morphological decomposition rules. For syntactic and semantic analysis we have obtained and tested a system of Interlisp programs from Bolt, Beranek and Newman which uses a compiled semantic augmented transition network grammar. We have created a dictionary of anatomical terms and are continually improving it. Currently, we are adding to it the features used by the BBN grammar. We still need to deepen the analysis of language embodied in our programs. We would like to move beyond the morphological and syntactic levels and attempt to capture the conceptual and intentional aspects of the language of histologists. We plan to investigate programs and approaches that other members of the SUMEX community have developed or use.

We have begun work on the computerized representation of other knowledge used by histologists, such as that regarding different culture media and tissue staining techniques. We are fortunate to have on this project Dr. Kahn, who is active in the Tissue Culture Association and a member of its Committee on Chemically Defined Media. This committee has been working for the past several years to compile and index, using computer files and programs, the components of and the bibliographic references for synthetic culture media. We plan to incorporate their compilation into the database that we are currently designing. We then hope to create a similar database for tissue stains.

We have not done a significant amount of work on the image analysis side of the project this year. We have spent most of our time waiting for new hardware to help us in this regard. The Anatomy Department has recently purchased a Quantimet 720 Image Processor which contains hardware to digitize microscope slides and perform basic image analysis functions. It also includes a PDP-11 computer and peripherals for storing images and further processing them. We plan to develop image analysis programs on this machine and at a later time interface them with our SUMEX programs. We also hope to have available soon a very fast image analysis computer that the Environmental Research Institute of Michigan is completing. We tested successfully on its prototype, programs for detecting prostate alveoli from digitized microscope slides.

E. Funding Support Status

Funding for the AI related portions of this project has been supplied solely by the University of Michigan, through the Mental Health Research Institute. An application for a grant from NIH was recently rejected. A revised application may be prepared, if we feel we can meet the objections of the NIH Study Section, which felt that the "goal is worthy and the computer areas that are proposed for implementation are of great usefulness and worth pursuing," but that the project as outlined in the proposal is overly ambitious.

II. INTERACTIONS WITH THE SUMEX-AIM RESOURCE

A. Collaborations, Interactions, and Sharing

The interaction and software sharing tools provided by SUMEX, such as SNDMSG, FTP, and protection mechanisms, were very useful to us in obtaining the ATN grammar compiler system from Bolt, Beranek and Newman. We found SNDMSG to be better than the mail or even the telephone for quick communication about our desires and problems, especially since our primary contact at BBN, Dr. Richard Burton, travels frequently but is almost always near a terminal. We used FTP to transfer the files from the BBN TENEX computer site to SUMEX. Dr. Burton was able to actually test the programs on the SUMEX system to ensure their proper functioning. We were surprised at the relatively little effort it took to transfer a system of this size and to obtain a working version. In addition, the ease with which we can communicate should enable us to continue our software sharing when new versions of these programs are developed

We have also had interactions with the SUMEX staff regarding minor Interlisp and TYMNET problems. We have found these staff consultants to be very helpful.

B. Critique of Resource Management

We are very pleased with the overall reliability and operation of the SUMEX system. Of the three years that we have been SUMEX users, this year has been the best in terms of reliability and response time. Now that the problem with dropped TYMNET characters has been solved and there is available a 120 cps TYMNET line, we find the editing and formatting facilities of SOS, PUB, SNDMSG, and TYPE to aid us in our formal and informal writing and to increase our productivity of reports and correspondence. We continue to find the SUMEX staff friendly and helpful. Our only criticism is a shortage of filespace. For example, we have one file that by itself takes approximately 100% of our allotted space.

III. RESEARCH PLANS (8/78 - 7/81)

A. Long Range Project Goals and Plans

Our long range technical goals have been described in the above summary. There are obviously many things to be done and some selection will need to be made. In the absence of expanded resources we will for the present concentrate on refining the language analysis systems already programmed, since these will eventually need to be extended to encompass a larger vocabulary and richer syntax. We will also work on the problem of converting images of tissue section into machine storable form, attempting to make this process more rapid and economically viable. We will also, as resources permit, develop the databases of media and of stains, both by enlarging them and by attempting to develop more useful structures for them. We feel that these projects will have some payoff even without the existence of the entire system envisioned.

B. Continuation of SUMEX Use

We have been using a rather small portion of the SUMEX resource, but have found it to be extremely valuable in our work; in fact, it has been absolutely essential. We wish to continue at approximately the same level until additional funding is obtained for the project. Our only serious limitation is in our allotment of disk space. Now that we have developed some software and compiled the beginnings of a dictionary our file space is quite cramped, and we would like to request a modest increase in our allocation.

Our programs are written in Interlisp, and SUMEX is the only facility to which we have access which supports this language. Perhaps the most important aspect of our association, however, is the opportunity SUMEX provides us for contact and interaction with other members of the AI community around the country.

We feel that we have made progress on an interesting problem which is highly relevant to the SUMEX-AIM mission. The descriptions of this work have been given in earlier sections of this report. Our hope is that this work will justify our continued access to the SUMEX facility.

C. Needs and Plans for other Computational Resources.

To develop our tissue culture model in its entirety, we need hardware for image processing. We believe that the Quantimet 720 computer recently purchased by the department of our medical collaborators can fulfill our needs in this regard. One use of the Quantimet which we foresee, is as a satellite computer to SUMEX, providing our SUMEX programs with pre-processed image data.

D. Recommendations for Future Community and Resource Development

Our basic recommendation to the SUMEX staff is to continue to provide high quality service as it has been doing. However, to guide management in selecting enhancements may we suggest the following.

1) The number of different models of display screens which the TV program can handle should be expanded so that more users can take advantage of screen editing.

2) To facilitate further software sharing the bulletin board system should be modified so that SUMEX users can describe software they have developed which might be of use to other projects. These descriptions should include for each program a person who can be contacted for more information. The current SUMEX system does a good job in sharing utility programs and documenting complete systems (such as MYCIN), but not programs that are more application oriented. While Stanford users are often able to obtain needed information by word of mouth, network users are not.

3) Any means for enhancing interactions among network users would be a great boon. For example, LINKing may not be terribly useful to Stanford users, but it could be more valuable to network users if it were more a matter of routine.

4) Along the lines of creating a greater feeling of community, perhaps an AI news service would be useful. This would inform users of current news of interest to the AI community as a whole, not just SUMEX-AIM news. Announcements of publications of important books or reviews, chess matches, government funding decisions, and so forth would be candidate items. Basically, news of the sort found in the SIGART Newsletter would be appropriate; this service would fill the gaps between Newsletters. It would not be necessary to aim at exhaustive coverage; any contributions at all would be useful.

5) Now that the bulletin board system has be in use for some time, perhaps it would be appropriate to review its design. We find it too cumbersome, and so tend to use it less than we might.

4.4 PILOT STANFORD PROJECTS

The following are descriptions of the informal pilot projects currently using the Stanford portion of the SUMEX-AIM resource pending funding, and full review and authorization.

4.4.1 GENETICS APPLICATIONS PROJECT

Computer Science Applications in Genetics

Prof. L. L. Cavalli-Sforza
Department of Genetics
Stanford University School of Medicine

I. SUMMARY OF RESEARCH PROGRAM

I.A. TECHNICAL GOALS

We are interested in understanding the role of diseases in shaping the geographic distribution of human genes. We observe a great deal of geographic variation but are still almost completely in the dark about its causes. Some clear examples have been given in the past that have linked the geographic distribution of a gene with that of a specific disease: sickle cell anemia is the best known example. In other cases the correlation was with a specific custom (e.g. lactose tolerance and the dietary use of untransformed milk).

We have been interested at the beginning in building gene frequency maps and will report here on the state of the art and the first applications of the technique.

I.B. MEDICAL RELEVANCE

Certain genetic diseases show conspicuous geographic and ethnic variation (examples: cystic fibrosis and phenylketonuria are highest in Caucasians; Tay-Sachs is found almost only among Ashkenazi Jews; gluten intolerance reaches frequencies of 1/300 in W. Ireland, and so on) - the causes are usually unknown. It could be that chance (random genetic drift) is involved.

When specific customs, infectious diseases, or pollutants can be pinned down as causal antecedents, there is usually room for preventive action. Specific interest is today given to the possibility that certain diseases may be determined by the interaction of specific stimuli on specific genotypes (e.g. emphysema in certain antitrypsin alleles; heart condition and arteriosclerosis in certain disturbances of lipid metabolism). Many of such conditions may be detectable by the study of geographic distributions and the correlations of genes and diseases.

I.C. PROGRESS SUMMARY

By far the major project this year has been the construction of GENE FREQUENCY MAPS. Alberto Piazza and Paolo Menozzi have collaborated on this project. There are some programs available which compute isopleths, but the underlying principles are either insufficiently specified, or do not correspond well enough with our needs. The data from which curves are to be built, gene

frequencies, are affected by sampling error. Thus a process of surface fitting is necessary; a simple interpolation, e.g. by splines is not adequate. Moreover, attempts in different directions in the last year have convinced us that one cannot use the same fitting method for the whole surfaces but the BEST LOCAL STRATEGY OF FITTING depends on the pattern of data actually available in the neighborhood of each point to be interpolated. This is critical especially for marginal regions. We have therefore developed a program in which a set of rules is given for the strategy to be chosen, and one out of a few different strategies is selected for each point to be interpolated in a network of an appropriate mesh. The mesh is chosen on the basis of computer time vs. desired resolution. Ordinarily, more points are necessary for drawing good isopleths than these fitted points that from the nodes of a relatively rough network. We have found in earlier work that drawing isopleths followed by shading the areas defined by isopleths is possible, but inefficient. Therefore, we have at the moment avoided computing the isopleths, and use instead splines under tension to generate a mesh of the desired high degree of resolution and which goes through all the previously calculated modes of the rougher network. The fine network generated by splines can be used for instance for output on TV screen, and this method has so far been giving us the most satisfactory results.

The maps we generate are, we believe, more satisfactory, and certainly more objective and far less time consuming than hand constructed maps. This, however, is not the major scientific benefit to be obtained from map construction by computer. We think the following is, instead the major result of interest. For the study of some factors of evolution, in particular migration, it is essential to appropriately average data over as many genes as possible. We take as weighted averages of gene frequencies the linear functions called "principal components" (corresponding to leading eigenvectors of the dispersion matrix). These have the advantages of maximizing the variation between populations. Plotting isopleths of the leading principal component, instead of gene frequencies, one can summarize the information contained in all gene frequencies known. In this way we have been able to show very recently that migrations as old as that occurring 9000 to 5000 years ago, with the diffusion of farmers from the Near East are still visible - and in fact are the dominant component - in the gene frequency map of Europe. Similar conclusions can be obtained from gene frequency maps of single genes but they are far less clear and convincing, since only a few genes show the expected gradients in a clear fashion, and only by appropriately summing the evidence from many genes does a clear picture emerge.

For this purpose of computing principal components however, one needs to have compact matrices of gene frequencies x populations with no missing items. Most real data are very far from this ideal condition. Having developed a satisfactory program to make gene frequency maps, we could then proceed as follows: make for each gene a map; sample for all genes the map values at a specified network points; compute principal components from this sample of interpolated values. The procedure was validated in a case in which we could obtain principal components independently, (on a real data matrix without missing items). We were thus able to generate maps of principal components of gene frequencies in spite of the extreme incompleteness of the raw data.

The results we have obtained by this new technique have gone beyond our expectation. We find that the amount of information we can synthesize by principal components tends around values of 30% for the first, 20% for the second

and 10% for the third. Thus we can synthesize 60% of the information by plotting the first three principal components in color (on a color TV screen produced by Grinnell of Palo Alto). This gives a very clear picture of the similarities between populations, as they can be gauged by synthesizing the maximum amount of information which can be stored in 3 dimensions (see figure 1).

II. INTERACTION WITH THE SUMEX AIM RESOURCE

A) We have asked the collaboration of all laboratories involved in HLA research so as to get data - published and unpublished - for a gene frequency bank. These data, as received, are stored in files. We plan to publish summaries at intervals (every 1-2 years) for distribution to collaborating laboratories, and publish them. The data will also be important for keeping up-to-date maps of HLA.

B), C) Nothing important.

III. RESEARCH PLANS

IV.A. LONG TERM PROJECT GOALS AND PLANS

Recently a big volume collecting data on gene frequencies for the most important human genes (excluding HLA and Gm) has been published by Mourant and others. This supersedes earlier attempts I had started to collate the same material, which had generated a less complete collection. Mourant's work has been that of a lifetime, but there were fears he might not be able to finish and publish his work. The second edition of his book which finally appeared in 1976 has been overdue for at least 6 or 7 years. The data thus available, from over 4000 original papers, deserve a full analysis. We have, I believe, made a good start to test them for one evolutionary factor, migration. This is a preliminary step of interest per se, in that it allows to reconstruct, for instance, population history; it is also an important basis for the study of other evolutionary factors. Having first eliminated the effect of migrations one can look at residual effects due to natural selection (and therefore diseases) and drift. Having built satisfactory maps of gene frequencies one can find selective responses to climatic conditions by correlating gene frequencies and climatic data. We have had remarkable results using anthropometrics instead of gene frequencies in a related project; an initial trial with gene frequencies has been moderately successful. We can also look at the correlation of the geographic distribution of diseases, as available from WHO statistical summaries and other sources, with that of single genes. There are also atlases of cultural data (e.g., Murdock's Ethnographic atlases) and an archive of languages which is computerized (c/o Stanford Department of Linguistics) which can be used to supply maps of cultural influences and which may be directly related to gene exchange, given that cultural and genetic migration are some time simultaneous, and that culture is another aspect of the environment in which we live. These are just indications of the lines of research we want to follow, on a long term basis for understanding more about factors of genetic differentiation and of environmental factors (including certain cultural ones) and their impact on human diversities with special consideration to health factors. Naturally, we are aware that the

finding of a correlation is only a first step in the detection of causal antecedence, and there are many ways in which correlations can fool us if accepted initially or without further work at other levels. The study of correlations is to be thought, however, as a screening method for detecting areas when further research may be useful. In the way we practice it, moreover, we have several ways of showing that a correlation is spurious. If we look at the correlation of a gene frequency with say the incidence of an infectious disease, (or a climatic variable), we expect the correlation to be found not only at a global level, but even more clearly in different world subareas. Moreover, our work on migration and other evolutionary factors allows us to eliminate other causes of variation of gene frequencies, namely all those that are common to many genes and not idiosyncratic to some. The latter is instead likely to be true of most instances of natural selection determined by disease.

On a shorter-term basis, we are interested in first improving our present program of map building. They already represent a first attempt at automation of a complex series of decisions in map building. We can probably improve greatly by recourse to Dendral. If we are successful in making the building of maps more precise and also more efficient we can hope to enlarge the range of problems which can be considered. From a technical point of view we can improve maps by simultaneously fitting gene frequencies for all alleles at a locus (whenever there are more than two alleles) because we can by simultaneous fitting take account of the restriction that all allele frequencies must sum to one. At the moment we do not consider this constraint. A more important program is that of studying drift by simultaneous fitting of several gene frequency maps. Populations which are of smaller size and are therefore expected to have more drift will show greater local variation. This is expected to show as local anomalies of the gene frequency map FOR SEVERAL GENES. A simultaneous fitting is therefore the simplest way of looking for such anomalies.

It is however more satisfactory to entrust the computer with as many as possible of the decisions to be taken, and the interaction would if anything be used as a temporary step for understanding more about the problems which arise and whose existence we may be at the moment unable even to anticipate. We are now trying to secure the help of a computer scientist to work on a more permanent basis on this problem. We thus hope to be able to make use of more sophisticated techniques, e.g. Dendral, to solve our problems.

We have been using in the past Sumex also for other projects; one of these, perhaps the biggest is Dr. Ammerman's storage-analysis-retrieval system of neolithic Calabrian data. But at present the load we have as compared with the number of pages available is such that we can only do one type of thing at a time, storing everything else on tape in the meanwhile. I had to ask Dr. Ammerman to find room elsewhere, in Sumex or another computer, even though I am vitally interested in his results and am planning to collaborate with him on their analysis when the storage will be finished. Other projects going on intermittently involve techniques of multivariate analysis. We are especially interested in developing new techniques for answering specific problems, for instance 1) the simultaneous fitting of means, in addition to variances and covariances. At the moment, only the latter are considered. One result we have had is that parameters derived from variances/covariances only do NOT fit means (e.g., the means of behavioral characteristics, like IQ, for adopted children given those of parents). This may be due to an insufficiency of the model, or of

the present fitting process, or of both; 2) the search for new synthetic ways of expressing multivariate data which maximize specific quantities: e.g. a function maximizing the correlation between adopted child and biological parent (which will express genetic factors) and one maximizing the correlation between adopted child and adoptive parent (which will express cultural factors).

I have been using multivariate analysis extensively and am convinced that, if properly used, it is the major help that artificial intelligence can give when complexity is due to the multiplicity of traits being considered. There are probably few other people interested in multivariate analysis in the Sumex community, since all that is available is a Factor Analysis program in SPSS and there are no programs for handling matrices (inversion and especially spectral analysis). Although there has been considerable increase of interest (especially among psychologists) multivariate analysis techniques remain relatively unused. I find that several of my problems had to await a really satisfactory solution for a long time until I tried classical techniques of multivariate analysis, the "principal components" we used for building synthetic gene frequency maps. Basically, multivariate analysis simplified problems by reducing the dimensionality of the data. If it were limited to this, one might think of it as a "mechanical" kind of artificial intelligence. There is, however, a dynamic aspect to be developed, in that there are many options and the choice of one could be made a more intelligent and automatic process than it is now.

There are many things we would have been unable to do without Sumex, but we have also been under constant shortage of space. I doubt we will be able to extend our work on maps if we are unable to obtain more space.

III.C. NEEDS AND PLANS FOR OTHER COMPUTATIONAL RESOURCES

We have a couple of projects at the planning stage and for which we many need more access to a computer. As yet I am not aware whether they will be of A.I. stype or involve more conventional computer usage. Accordingly, in the financing of the project (now been submitted to NSF) we ask for a small size minicomputer. Another resource of interest is a color TV screen, of the type we have used by kind permission of Prof. R. Lyon of the Earth Sciences Department. We are considering the possibilities of acquiring one, if we can find the financing.

4.4.2 QUANTUM CHEMICAL INVESTIGATIONS

Theoretical Investigations of Heme Proteins, Opiate Narcotics, Chemical Carcinogens, Drug Metabolism, DNA Structure and Intercalating Drugs

Dr. Gilda Loew
Molecular Theory Laboratory
Department of Genetics
Stanford University

I. SUMMARY OF RESEARCH PROGRAM

The major theme of research in the Molecular Theory Group is the application of the techniques of theoretical chemistry to a variety of biomedical problems. To this end a number of large scale computer programs which embody these techniques are utilized to characterize the electronic structure, conformation, interactions with appropriate receptors, and electromagnetic properties of biologically important molecules. The SUMEX-AIM resource is used to characterize starting geometries of appropriate molecular systems used as input to the large scale programs and to calculate electromagnetic properties from the output of such programs. It is also used for rapid energy calculations of opiate peptide conformations. This judicious use of the facility thus enhances the productivity and efficiency of all of of our research projects. A brief description and status of specific areas of research are given below.

A. Structure Activity Studies of Opiate Narcotics

Many classes of opiates have been studied including the newly discovered endogenous peptide opiates. Similarities and differences among these classes have been delineated and molecular properties related to agonist/antagonist potency ratios identified and characterized. (See references 1-3 for the latest of 11 publications in this area). This work is in its fourth year of support from the National Institute of Drug Abuse. Collaboration with medicinal chemists and clinical pharmacologists interested in drug design is under way.

B. Drug Metabolism and Activation of Chemical Carcinogens by Cytochrome P450

Models of the active site of this ubiquitous metabolizing enzyme in its biologically active state have been made and electromagnetic properties predicted. In addition, structure activity studies of classes of substrates, specifically general anesthetics and polycyclic aromatic carcinogens, have been made which relate molecular reactivity to substrate efficacy and metabolite distribution (see references 4-9). This work has been modestly supported in the past by NSF and ACS and is currently being supported by a new grant from NIH (GM) and a one year contract from NCI.

C. Chemical Carcinogens

The goal of this study is to help identify and calculate properties of parent compounds related to their carcinogenic potency and hopefully to ultimately predict which members of given classes of chemical carcinogens will be active. Models for metabolic activation, mode of action and interaction with DNA are part of this project which is funded by a one year contract from NCI (same as above).

D. Structure Function and Electromagnetic Properties of Heme Proteins and Related Metal Organic Complexes

Using a reasonable model as input, large scale molecular orbital programs are used to calculate the electronic structure and conformation of the active site of heme proteins and the mode of binding of a number of biologically relevant ligands such as CO, O2 and CN-. Using the calculated electronic structures and conformations, a set of auxiliary programs are used to calculate measurable electromagnetic properties such as quadrupole splitting observed in Mossbauer resonance spectra, g values and hyperfine splittings in electron spin resonance spectra, and magnetic moments. This close connection between observables and basic molecular structure enhances the usefulness of experimental data in inferring such fundamental molecular properties as the nature of metal-ligand binding and how small changes in conformation at the active site affect biological function (references 10-12). This work has had modest continual support from NSF for the past twelve years

E. DNA Structure and Interaction with Intercalating Molecules

This work is divided into two parts: Investigations of the various modes that DNA can "kink" consistent with known chromatin structure and the origin of the specificity in the binding of ethidium-type intercalators into DNA components (references 13-15). This work is not currently funded.

II. INTERACTION WITH SUMEX-AIM RESOURCE

The SUMEX-AIM resource is used by the Molecular Theory Laboratory on a limited basis which allows optimum efficiency on almost all of our projects. Specifically we have used it for:

A. Interactive Data Preparation and Input for all Projects

The procedure for the characterization of heme proteins initially requires a determination of a model for the active site. The model has to be large enough to realistically describe interactions that take place at the active site but not so large that it becomes economically infeasible to perform a calculation on the molecule. Once a general model is chosen, the specific geometry for the molecule must be determined. With the aid of experimental crystal structures of related compounds, we use the SUMEX-AIM facility to interactively determine the coordinates of the best initial approximation to the geometry of the active site model. Specifically for heme proteins this involves determining the hole size of the porphin ring, the degree of nonplanarity of the porphin ring, the position of

the metal atom at the center of the active site, and the nature and geometry of the axial ligands bound to the metal.

For the study of the relative properties of a family of carcinogens as a function of substituent on a given parent compound, the SUMEX-AIM facility is used to interactively determine and plot the most reasonable geometries of the various substituents depending on the parent compound. Similarly, the geometries of DNA base-pairs and organometallic transition metal complexes are calculated using SUMEX-AIM.

B. Calculation of Electromagnetic Properties of Iron Containing Proteins and Related Organometallic Compounds

We are currently performing systematic studies of heme proteins including the metabolizing enzyme cytochrome P450. The electromagnetic properties of these proteins and of synthesized model compounds which mimic the observed behavior of the proteins have been well studied experimentally in a number of instances. SUMEX-AIM is used for the calculation of these one-electron properties.

The properties that are calculated include the electric field gradient at the iron nucleus, quadrupole splitting, isotropic and anisotropic hyperfine interaction, spin-orbit coupling and zero field splitting, g values and temperature dependent effective magnetic moments. The calculated values are compared directly to experimental results obtained from published Mossbauer resonance and electron spin resonance spectra. Such a comparison determines not only the reliability with which these properties can be calculated but also gives an indication of the ability of the model of the iron active site to mimic the actual environment found in a particular compound or iron containing protein.

The major input to these properties programs is a description of the electron distribution of the compound under consideration. This description is obtained using a semiempirical molecular orbital method employing the iterative extended Huckel procedure. Such a calculation requires up to 660K core and is performed elsewhere. When the calculated electron distribution yields a set of calculated properties in agreement with observation, we have increased faith in the description of the model of the active site and can carry the model one step further to make qualitative inferences about certain properties relevant to the biological function of the compound.

C. Conformational Study of Pentapeptides

In a completely different context, we have used SUMEX-AIM to calculate the conformation of pentapeptides (enkephalins) which have been recently found to be endogenous opiates. The aim of this study was to determine in what way, if any, they can mimic the structure of prototype opiates such as morphine and meperidine. For this work, we have used a protein conformation program with empirical interaction potentials. Quantum mechanical conformation calculations of the same peptides have been performed by us elsewhere and the results of the two methods have been compared (reference 2).

III. RESEARCH PLANS (8/78-7/81)

We plan to continue research in the same general areas as our current projects: Systematic studies of iron-containing proteins; structure activity studies of opiates and other drugs; drug metabolism and activation of chemical carcinogens; structure activity studies of classes of chemical carcinogens; structure of DNA and binding of DNA components to small drugs and carcinogens; and studies of peptide conformation.

Specific judicial use of SUMEX-AIM is planned in connection with most of these projects in the same spirit as we have described for our on-going projects:

1. We plan to use SUMEX-AIM to continue to characterize reasonable molecular geometries to use as input to our large scale computer programs which calculate molecular properties and also intermolecular interactions (drug/receptor, enzyme/substrate). The flexible operating mode of SUMEX-AIM is important to the efficient and accurate preparation of input before costly time consuming production runs are submitted at another facility.

2. We plan to use SUMEX-AIM to continue our systematic study of heme proteins, specifically in the calculation of electromagnetic properties. These calculations will help to characterize the four stable states of the metabolizing heme enzyme cytochrome P450 and to study the relationship between conformation and function of other classes of heme proteins (electron and oxygen transfer proteins).

3. We plan to use SUMEX-AIM for a new study of peptide conformations using the experience gained from the investigation of peptide opiates. Specifically, in collaboration with G. Fassman we plan to investigate specific correlations between primary sequence and conformation of proteins and to do limited structure activity studies of some peptide hormones.

4. We plan to use SUMEX-AIM to help decide plausible geometries for complexes formed by the active form of chemical carcinogens with DNA components.

SUMEX-AIM has become an essential component of our research procedure. The continued use of SUMEX-AIM is central to the efficient implementation of our research plans. We are grateful for past support and feel justified in our request for continued support.

REFERENCES

1. Loew, G. H., Berkowitz, D. S., Quantum Chemical Studies of N-substituent Variation in the Oxymorphone Series of Opiate Narcotics, J. Med. Chem., 21, 101 (1978).

2. Loew, G. H. and Burt, S. K., An Energy Conformation Study of the Resemblance of Met-Enkephalin and D-Ala2 Met-Enkephalin to Rigid Opiates Using Empirical and Quantum Mechanical Methods. Proc. Natl. Acad. Sci. 75, 17, (1978).

3. DeGraw, J. I., Lawson, J. A., Crase, J. L., Johnson, H. L., Ellis, M., Uyeno, E. T., Loew, G. H., and Berkowitz, D. S., Analgesics. I. Synthesis and Analgesic Properties of N-Sec-Alkyl and N-Tert-Alkylnormorphines, J. Med. Chem. (1978), in press.

4. Hjelmeland, L., Loew, G. H., The Electronic Structure of Peracids: Functional Models for Cytochrome P450, Tetrahedron, 33, 1029 (1977).

5. Loew, G. H., Kert, C., Hjelmeland, L., Kirchner, R. F., Active Site Models for Horseradish Peroxidase Complex I and a Cytochrome P450 Analogue, J. Amer. Chem. Soc., 99 (10), 3534 (1977).

6. Hjelmeland, L., Loew, G. H., The Electronic Structure for Chromyl Chloride: A Functional Model for Cytochrome P450, J. Am. Chem. Soc., 99 (10), 3514 (1977).

7. Loew, G. H., Hjelmeland, L., Kirchner, R. F., Models for the Enzymatically Active State of Cytochrome P450, Int. J. of Quant. Chem., QBS, 4, 225 (1977).

8. Loew, G. H., Wong, J., Phillips, J., Hjelmeland, L., Pack, G., Quantum Chemical Studies of the Metabolism of Seven Benzo(a)pyrene, Can. Biochem. Biophys., in press (1978).

9. Loew, G. H., Phillips, J., Wong, J., Hjelmeland, L., Pack, G., Quantum Chemical Studies of the Metabolism of Seven Polycyclic Aromatic Hydrocarbons, Can. Biochem. Biopys., in press (1978).

10. Kirchner, R. F. and Loew, G. H., Semiempirical Calculations of Model Oxyheme: Variation of Calculated Electromagnetic Properties with Electronic Configuration and Oxygen Geometry, J. Am. Chem. Soc., 99, 4639 (1977).

11. Loew, G. H. and Kirchner, R. F., Semiempirical Calculations of Model Deoxyheme. Variation of Calculated Electromagnetic Properties with Electronic Configuration and Distance of Iron form the Plane, Biophys. J., 22, 000 (1978).

12. Loew, G. H. and Kirchner, R. F., Binding of O2, NO, and CO to Model Active Sites in Ferrous Heme Proteins: Ligand Geometry, Electronic Structure and Quadrupole Splittings, Int. J. Quan. Chem. QBS, in press (1978).

13. Pack, G. R., Muskavitch, M. A., Loew, G. H., Kinked DNA: Energetics and Conditions Favoring its Formation, Biochim. Biophys. Acta, 478, 9 (1977).

14. Pack, G. R.and Loew, G. H., The Origins of Sequence Specificity of Ethidium Nucleic Acid Intercalation, Int. J. Quant. Chem. QBS, 4, 87 (1977).

15. Pack, G. R. and Loew, G. H., Origins of the Specificity in the Interaction of Ethidium into Nucleic Acids: A Theoretical Analysis, Biochim. Biophys. Acta, in press (1978).

Appendix I

AIHANDBOOK OUTLINE

E. A. Feigenbaum
A. Barr
Computer Science Department
Stanford University

This is a tentative outline of the Handbook. Articles in the first eight Chapters are expected to appear in the first volume. A list of the articles in each Chapter is appended.

I. Introduction
II. Heuristic Search
III. AI Languages
IV. Representation of Knowledge

V. Natural Language Understanding
VI. Speech Understanding
VII. Applications-oriented AI Research
VIII. Automatic Programming

IX. Theorem Proving
X. Vision
XI. Robotics

XII. Information Processing Psychology
XIII. Learning and Inductive Inference
XIV. Problem Solving, Reasoning, and Planning

I. INTRODUCTION
- A. The AI Handbook (intent, audience, style, use, outline)
- B. Overview of AI
- C. History of AI
- D. Philosophy of AI
- E. AI and Society
- F. An Introduction to the Literature in the field

II. Heuristic Search
- A. Overview
- B. Problem representation
  1. Overview
  2. State space representation
  3. Problem reduction representation
  4. Game trees
- C. Search
  1. Blind state space search

2. Blind search of an and/or tree
3. Heuristic search in problem-solving
    a. Basic concepts in Heuristic Search
    b. A*: optimal search for an optimal solution
    c. Relaxing the optimality requirement
    d. Bidirectional search
    e. Heuristic search of an and/or graph
    f. Hill climbing
4. Game tree search
    a. Minimax
    b. Alpha-beta

D. Example Programs
1. Logic Theorist
2. GPS
3. Gelernter - geometry
4. Slagle & Moses - Integration
5. STRIPS
6. ABSTRIPS

III. AI Languages
A. Overview of AI Languages (Historical)
B. Comparison of AI Languages
C. Early list-processing languages (SLIP, IPL, SNOBOL)
D. Current languages/systems
1. LISP, the basic idea, INTERLISP
2. SAIL/LEAP
3. POP-2, POP-10
4. QLISP (mention QA4)
5. PLANNER
6. CONNIVER
7. Object-oriented languages (ACTORS, SMALLTALK, SIMULA)
8. FUZZY (LeFaivreaRutgers)
9. QA3/PROLOGUE
10. PC programming languages

IV. Representation of Knowledge
A. Survey of representation techniques
B. Issues and problems in representation theory
C. Representation Schemes
1. Predicate calculus
2. Semantic nets (Quillian, LNR, Hendrix)
3. Production systems
4. Procedural representations (SHRDLU, actors, demons)
5. Semantic primitives, Componential analysis
(Fillmore, Schank, Wilks)
6. Direct (Analogical) representations
7. Higher Level Knowledge Structures
    a. Frames, Scripts, The basic idea
    (Bartlett, Minsky, Abelson)
    b. KRL-0, MERLIN
    c. Others: FRL, OWL, Toronto

D. Discussion and Comparison of Representation Schemes

V. Natural Language
- A. Overview - History & Issues
- B. Grammars
  1. Review of formal grammars
  2. Transformational grammars
  3. Systemic grammars
  4. Case Grammars
- C. Parsing techniques
  1. Overview of parsing techniques
  2. Augmented transition nets, Woods
  3. CHARTS - GSP
- D. Text Generating systems
- E. Machine Translation
  1. Overview & history
  2. Wilks' machine translation work
- F. Natural Language Processing Systems
  1. Early NL systems (SAD-SAM through ELIZA)
  2. PARRY
  3. MARGIE
  4. LUNAR
  5. SHRDLU, Winograd

VI. SPEECH UNDERSTANDING SYSTEMS
- A. Overview (Includes a mention of ac. proc., blackboard)
- B. Design Considerations for Speech Systems
- C. The Early ARPA speech systems
  1. DRAGON
  2. HEARSAY I
  3. SPEECHLIS
- D. Recent Speech Systems
  1. HARPY
  2. HEARSAY II
  3. HWIM
  4. SRI-SDC System

VII. Applications-oriented AI research
- A. Overview of AOAIR
- B. CHEMISTRY
  1. Mass Spectrometry (DENDRAL, CONGEN)
  2. Organic Synthesis (Wipke@UCSC)
- C. MEDICINE
  1. Overview
  2. MYCIN
  3. Glaucoma
     a. CASNET,
     b. IRIS
  4. DIALOG, INTERNIST II
  5. PRESENT ILLNESS (MIT)
  6. DIGITALIS (MIT)
  7. TEIRESIAS
- D. MATHEMATICS
  1. REDUCE

2. MACSYMA
3. AM

E. EDUCATION
1. Overview
2. SCHOLAR
3. SOPHIE
4. WEST, BUGGY
5. COACH

F. MSC.
1. SRI Comp. Based Consultant (PROSPECTOR)
2. RAND-RITA
3. AI applications to Information Retrieval
4. SU/X
5. Management applications

VIII. Automatic Programming
A. Automatic Programming Overview
B. Program Specification
C. High-level Program Model Construction
C. Program Synthesis
1. Overview
2. Techniques (Traces, Examples. Natural Language, TP)
a. Traces
b. Examples
c. Natural Language
d. Theorem Proving
D. Program optimization techniques
E. Programmer's aids
F. Program verification (see Article IXD5)
G. Integrated AP Systems

IX. THEOREM PROVING
A. Overview
B. Predicate Calculus
C. Resolution Theorem Proving
1. Basic resolution method
2. Syntactic ordering strategies
3. Semantic & syntactic refinement
D. Non-resolution theorem proving
0. Overview
1. Natural deduction
2. Boyer-Moore
3. LCF
E. Uses of theorem proving
1. Use in question answering
[2. Use in problem solving]
3. Theorem Proving languages
4. Man-machine theorem proving
5. In Automatic Programming
F. Proof checkers

X. VISION
- A. Overview
- B. Image-level processing
  1. Overview
  2. Edge Detection
  3. Texture
  4. Region growing
  5. Overview of Pattern Recognition
- C. Spatial-level processing
  1. Overview
  2. Stereo information
  3. Shading
  4. Motion
- D. Object-level Processing
  1. Overview
  2. Generalized cones and cylinders
- E. Scene level processing
- F. Vision systems
  1. Polyhedral or Blocks World Vision
     a. Overview
     b. COPYDEMO
     b. Guzman
     c. Falk
     d. Waltz
     e. Navatya
  2. Robot vision systems
  3. Perceptrons
  4. etc.

XI. ROBOTICS
- A. Overview
- B. Robot Planning and Problem Solving
- C. Arms
- D. Present Day Industrial Robots
- E. Robotics Programming Languages

XII. Information Processing Psychology
- A. Overview
- B. Memory Models
  1. Overview
  2. EPAM
  3. Semantic Net Models
     a. Quillian & Collins
     b. HAM-ACT (Anderson & Bower)
     c. LNR ASNs
  4. Production Systems a Memory Models (Newell, Moran, ACT)
  5. Higher level structures (Schemas, scripts & Frames)
- C. Human Problem Solving
- D. Behavioral Modeling
  1. Belief Systems
  2. PARRY
  3. Conversational Postulates (Grice, TW)

4. Politics, Abelson R. and J. Carbonell, Jr.,

XIII. Learning and Inductive Inference
- A. Overview
- B. Simple Inductive Tasks
  1. Sequence Extrapolation
  2. Grammatical Inference
- C. Pattern Recognition
  1. Character Recognition
  2. Other (e.g. Speech)
- D. Learning Rules and Strategies of Games
  1. Formal Analysis
  2. Individual Examples of Games-learning programs
- E. Single Concept Formation
- F. Multiple Concept Formation: Structuring a Domain (AM, Meta-DENDRAL)
- G. Interactive Cumulation of Knowledge (TEIRESIAS)

XIV. Problem Solving, Planning & Reasoning by Analogy
- A. Overview of Problems Solving
- B. Planning
  1. Overview (pointers to discussions in Search, Robotics, AI Languages)
  2. STRIPS (see IID5)
  3. ABSTRIPS (see IID6)
  4. NOAH
  5. HACKER
  6. INTERPLAN (Tate)
  7. Rieger's inference engine
  8. BELIEVER (Schmidt, Sridharan)
  7. QA3 (see IXE1)
- C. Reasoning by Analogy
  1. Overview
  2. Evans's ANALOGY Program
  3. ZORBA
  4. Winston (see Learning)
- D. Constraint relaxation
  1. Waltz (see Vision)
  2. REF-ARF
- E. Game playing
  (This overview must point to work in search and discuss GP programs of various misc. sorts)

Appendix II

DIGITAL COMMUNICATIONS AND SCIENCE

Digital Communications and the Conduct of Science
THE NEW LITERACY

Joshua Lederberg, Senior Member, IEEE
Department of Genetics
Stanford University

Abstract: This essay is a personal perspective on the emergence of a new form of communication, optimistically called the 'eugram'. This form is based on the convergence of economical digital communications with computer aided facilities for file management, and protocols to facilitate the interconnection of users separated both in time and space. The eugram is contrasted with the telephone, with the latter's demands on instant availability and the subjugation of the user to an almost uninterruptible stream of data. The eugram is expected to increase the thoughtfulness of communication, the return of literacy in the efficient and precise use of language, and to enhance scientific discourse in many other ways.

Introduction

Computer communication networks provide new tools and opportunities for the scientific community to share scarce computer-based resources. They permit a new form of informal communication between scientists and often provide motivation and reward for timely sharing of research results. In addition computer-based support to large distributed segments of a scientific community is made possible via users and computers interconnected by computer controlled networks.

Today the most significant and useful form of computer communication is based on packet switched technology which has been reduced to practice in daily support of some portions of he scientific community.

Two key elements of this technology base are:

- Computer-based user-user message capability, i.e., electronic mail plus the computer-management of text data.

- Sharing in the development, refinement and use of large, complex computer knowledge-based systems particular to a segment of science, which would not otherwise be widely available.

This essay is written from the perspective of an enthusiastic USER of packet switched communications. The system itself is here regarded as a black box that accomplishes efficient transfer of digitally encoded information in near-real time among terminals that interface both to human users and to computer-manageable files. The economical integration of user, file,processor, and distance-indifferent communication link is the novel capability of what I shall call a EUGRAM system. EUGRAPHY thus embraces not only electronic despatch of mail but also a panoply of computer-augmented text handling tools and protocols. This account is informed by my experience over the last five years in the development of the SUMEX-AIM community for research in artificial intelligence related to biomedical science. However, it will be primarily concerned with the expected impact of, and needs for, the elaboration of EUGRAPHY in the conduct of scientific research generally over the next 25 years.

Conduct of Science: Computers and Communications

The claim of science to universal validity is supportable only by virtue of a strenuous commitment to global communication. In the spatial domain, the canon of publication insists upon public awareness and criticism of avowedly new knowledge. This enforces the reliability of empirical reports and assembles them into common models of a real world. In the temporal domain, the archiving and retrieval of information sustains the discipline of novelty -- assuring that we acknowledge, so as to be able to extend, the boundaries of 'human', i.e., universal knowledge.

The past twenty years have witnessed a growing self-consciousness about the structure of scientific activity, impelled in part by Malthusian concerns over the long term implications of a geometric increase at 0.25 db/yr: a ten-fold expansion over the 40-year typical career of the scientist. Much more has been written than implemented about means of helping scientists keep up with the "information-explosion". One must acknowledge the utility of recent introductions of literature-searching and alerting services, many of which crucially depend on computer support and EUGRAM-like communications. On the other hand, it will probably be the cost-explosion of print media for scientific publications [1] that proves to be a more immediately compelling motive for fundamental reexamination of our methods of scientific documentation and communication. Designs for solving these problems -- reviewed long since [2] -- must take into account that the media for communication also play a crucial role in quality control in science. The filtering procedures of the 'refereed journal' support the selection both of worthwhile reading, and of the workers whose established performance entitles them to the privileges of academic positions and social subsidy for their research.

Perhaps on account of these latter concerns, most of my colleagues in biomedical research would be loath to adopt many changes in the present system of print publication. In practice, frequent personal encounters [3] facilitated by grant funding, jet aircraft and invisible colleges [4,5] seem to play an increasingly important role in the exchange of information within scientific specialities, but without any systematic inquiry as to the costs, efficiency, and equitability of these modes. Nevertheless, no piece of work, no claim to priority, is authentically recorded until it has appeared in public print in a respectable refereed journal. The long-distance telephone surely has its role

also, but more for operational detail than serious intellectual discourse; and the use of the mails is as idiosyncratic as is the performance of the US Postal System, with the notable exception of the exchange of xerocopied preprints of forthcoming publications.

In the face of this inertia, one should be skeptical about the marketability of new systems like EUGRAPHY, regardless of their technical virtues. Indeed, scientists may be the last to adopt them on a comprehensive scale, except for demonstrations that may arise from a) computer science, b) research management, c) military requirements, d) the ever graver collapse of conventional mails, and e) business applications like EFTS. With respect to c), we of course owe a great deal to the ARPANET as showing the way, and with the potential for a spillover into civil technology perhaps comparable only to jet engine. The sheer economy of EUGRAPHY, and the diffusion of microprocessors and displays into the laboratory and into everyday life, are bound to force an encounter with the challenges of new systems despite the traditional conservatism of the scientific establishment (with respect to its own way of doing business, and its attention to change outside the academic discipline [6]). Nevertheless, the history of the medical and engineering sciences both show many instances where a reluctant marriage of theoria and praxis has engendered major enrichments of the basic sciences.

All the above notwithstanding, our own experience with EUGRAPHY at SUMEX-AIM has been extraordinarily good. Individual users of course rely upon it routinely for access to computer processing. More surprising was the utility of EUGRAPHY for research management, involving the exchange of texts even over relatively short distances -- offices down the corridor or in nearby buildings. This phenomenon has provoked introspections about EUGRAPHY as a qualitatively different method of interpersonal communication from conversation, the telephone, the handwritten memo, the dictated letter or the published report, and some speculations about the further evolution of EUGRAMs as part of scientific communication.

Comparing the EUGRAM with the Telephone [7,8]

When telephone usage is limited to a few calls per day, and the connecting parties are reliably locatable, the telephone may indeed fulfill its image of instant, spontaneous communication. In current practice, beleaguered by time zone shifts, lunch hours, conferences, and competing calls, the reality of phone usage is exemplified by the employment of secretaries to make and receive the calls. The very instantaneity of the phone connection generates a queueing problem that defeats the basic motive. In due course, the two-way conversation may disappear, to be replaced by messages stored on tape recorders. The information density of speech may be viewed as very low, or very high, depending on how much of the burden is carried by the text, how much by inflection, phrasing, and other personal qualities. It may be only with respect to communications that have high affective content that audio- can compete with digital channels, and to do this well may require better than the average channel quality than is now readily available between metropolitan centers. Even here, the enhancement of literary competence might go a long way to permit the EUGRAM to compete with the song.